Influence of molecular structures on graph neural network explainers' performance

Tim Stols, T.N.Stols@student.tudelft.nl

For the course CSE3000 Research Project – Supervisors: Dr. Megha Khosla, Dr. Jana Weber

1 Background

- Graph Neural Networks (GNNs) are used in molecule design for property prediction,
- Molecules are represented by atoms (nodes) and bonds (edges)[1].
- Although GNNs show promising performance [2], they lack explainability [3].
- Explainable AI (XAI) explainer models can highlight influential parts of the molecule for the output. (See Figure 1)
- Rathee et al. propose new metrics to evaluate these explainers on **faithfulness** (whether the explanation corresponds to the model's thinking) and **plausibility** (whether the explanation corresponds to human-provided truth) [4]. See Eq 1 & 2.
- GNN explainers for chemistry need to be assessed using these benchmarks, to improve the acceptance of the models and their explainers.
- Explainers need to be evaluated for performance on different subsets of the dataset, based on molecular properties.

2 Research Question

The main research question for this research project is: **What is the impact of different dataset properties on GNN explainer performance?**

Different splitting methods are used to separate the dataset into subsets:

- Containing or not containing benzene rings
- Containing or not containing halogens
- Below- or above-median molecular weight

BAGEL regimes of faithfulness and plausibility are evaluated on these subsets.

3 Methodology

Communicative Message Passing Neural Network (CMPNN) is used with Integrated Gradients (IG) explainer as baseline GNN model and explainer. These performed best in the evaluation by Rao et al. [3].

3MR dataset is used which tests for 3-membered rings in molecules. This concerns a binary classification task.

A Python notebook is implemented, in which a model and explainer are trained and evaluated with the BAGEL metrics.

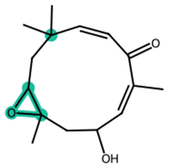


Figure 1: Binary explanation mask for 3-Hydroxy-1,5,9,9-tetramethyl-12-oxabicyclo[9.1.0]dodeca-4,7-dien-6-one

Equations 1 and 2 for faithfulness:

1. $Sufficiency = f(G) - f(G_E)$
2. $Comprehensiveness = f(G) - f(G \setminus G_E)$

where f(G) is the original prediction for the full graph G.
$f(G_E)$ is the model prediction for the explainer subgraph G_E.
$f(G \setminus G_E)$ is the model prediction for the graph without the explainer subgraph G_E.

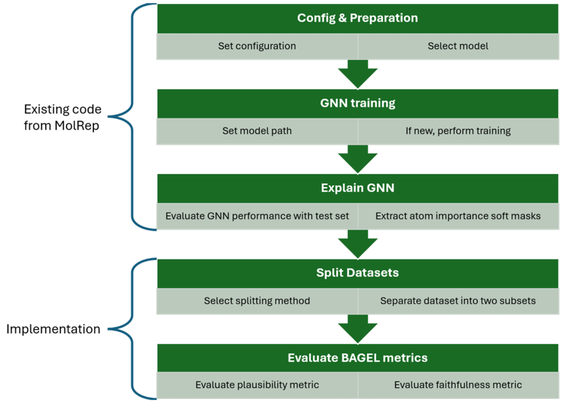


Figure 2: Flow chart of the functionality in the implemented Jupyter Notebook

4a Results - Faithfulness

The subsets with benzene, halogens, and high MW performed worse for both sufficiency and comprehensiveness with a low p-value below the threshold of 0.01.

These results show a correlation; however, the expected cause is a **correlation** with the **difference in positive rate** in the subsets.

This is shown in Table 1.

Table 1: Table showing the positive rate, sufficiency, and comprehensiveness measured on the complete dataset and experiment subsets. Difference columns (colored yellow) show the difference between the subset values and the value for the complete dataset. The yellow values in these rows match closely.

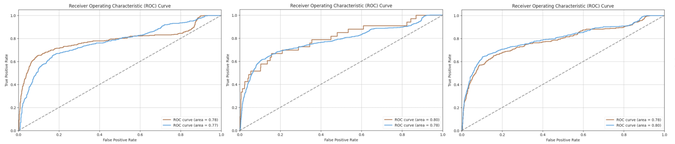
Subset	Positive rate	Difference from complete	Sufficiency score	Difference from complete	Compre-hensiveness score	Difference from complete ×2
Complete	0.4236	-	-0.08	-	0.26	-
Containing benzene	0.3062	0.1174	-0.19	0.11	0.2	0.12
Not containing benzene	0.6329	-0.2093	0.08	-0.16	0.36	-0.2
Containing halogens	0.1111	0.3125	-0.33	0.25	0.07	0.38
Not containing halogens	0.4885	-0.0649	-0.03	-0.05	0.29	-0.06
High MW	0.3175	0.1061	-0.1	0.02	0.23	0.06
Low MW	0.4757	-0.0521	-0.05	-0.03	0.3	-0.08

4b Results - Plausibility

Evaluation of the ROC curves when compared against ground truths, the actual 3-membered rings as hard masks.

All subsets show a similar Area Under the ROC curve to the complete dataset

Benzene: The **shape** of the curve **differs** Halogens: The curve is **not continuous**



Figures 4-6: ROC curves for the generated explanations compared against ground truth attributions. Figure 4 shows the curve for the subset containing benzene (orange) and not containing benzene (blue). Figure 5 shows the curve for the subset containing halogens (orange) and not containing halogens (blue). Figure 6 shows the curve for MW >= 297,034 (orange) and MW < 297,034 (blue).

5 Conclusions and Future Steps

Faithfulness:

- **Positive rate influences** comprehensiveness and sufficiency.

Subsets should be equal in the positive rate or

BAGEL metrics must be changed for positive and negative samples to achieve the same range.

Plausibility:

- The chosen splits **do not influence** the overall accuracy of the explainer.
- Benzene ring influences performance for **certain** explainer **thresholds**.

Future:

- This research gives a **first step** in the evaluation of GNN explainers using the BAGEL metrics.
- **More evaluation** is needed to understand the behavior of the explainers better and to apply them more fittingly.
- The experiment should be performed on **new datasets**, to test a wider range of molecules.

References

1. Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., Hsieh, C., & Hou, T. (2023). Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. Nature Communications, 14(1). https: //doi.org/10.1038/s41467-023-38192-3
2. Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., & Zheng, M. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of Medicinal Chemistry, 63(16), 8749â8760. https://doi.org/10.1021/acs.jmedchem.9b00959
3. Rao, J., Zheng, S., & Yang, Y. (2021). Quantitative evaluation of explainable graph neural networks for molecular property prediction.
4. Rathee, M., Funke, T., Anand, A., & Khosla, M. (2022). Bagel: A benchmark for assessing graph neural network explanations.